Probabilistic Forecasts of Bike-Sharing Systems for Journey Planning

Nicolas Gast
Inria
Univ. Grenoble Alpes
CNRS, LIG, F-38000
Grenoble, France
nicolas.gast@inria.fr

Guillaume Massonnet
Inria
Univ. Grenoble Alpes
CNRS, LIG, F-38000
Grenoble, France
guillaume.massonnet@inria.fr

Daniël Reijsbergen
LFCS
University of Edinburgh
dreijsbe@inf.ed.ac.uk

Mirco Tribastone
IMT - Institute for Advanced
Studies Lucca
mirco.tribastone@imtlucca.it

ABSTRACT

We study the problem of making forecasts about the future availability of bicycles in stations of a bike-sharing system (BSS). This is relevant in order to make recommendations guaranteeing that the probability that a user will be able to make a journey is sufficiently high. To do this we use probabilistic predictions obtained from a queuing theoretical time-inhomogeneous model of a BSS. The model is parametrized and successfully validated using historical data from the *Vélib'* BSS of the City of Paris.

We develop a critique of the standard root-mean-square-error (RMSE), commonly adopted in the bike-sharing research as an index of the prediction accuracy, because it does not account for the stochasticity inherent in the real system. Instead we introduce a new metric based on scoring rules. We evaluate the average score of our model against classical predictors used in the literature. We show that these are outperformed by our model for prediction horizons of up to a few hours. We also discuss that, in general, measuring the current number of available bikes is only relevant for prediction horizons of up to few hours.

1. INTRODUCTION

In recent years, bicycle sharing systems (BSS) have proven to be very successful in several major cities and are now spreading all across the world [4, 5]. Nowadays, there exist more than 700 such systems that operate on five continents. The benefits for cities are multiple: from a greener image due to more eco-friendly means of transportation to the reduction of traffic congestion, noise and air pollution, they provide an alternative to private motorized vehicles, especially for short-distance trips. From the user's perspective, they offer an affordable and efficient transport alternative with several benefits over the use of a personal bicycle with respect to maintenance, theft or storage issues.

A BSS is composed of a number of stations where a limited number of bikes can be parked. The dynamics is simple. A user arrives at a station to pick up a bike. The ride ends when she returns the bike to any station. User experience and provider revenue can be hampered when the origin station is empty, which forces the user to either resort to another means of transport or try to find an available bike in another station. Similarly, if the destination station is full, the user must either wait until one bike is picked up, or return the bike to another station with at least one parking spot available.

BSS providers publish, typically through their websites, live data about the availability of bikes and empty parking spots at BSS stations, as a means of helping users plan their journey. This has caught the attention of many researchers, who used this data to study user behavior and mobility patterns [3, 6, 14]. Some notable effort has been put on the development of predictive models for BSS, which can be calibrated using the real available datasets [9, 12, 13, 15]. Having a predictive model can benefit both the user and the BSS provider. The user can employ it as a journey advisor [15], for instance identifying likely origin/destination stations for which the trip can be successfully made. BSS providers can run the model to discover usage trends, and proactively adopt countermeasures to potentially adverse situations, such as performing a rebalancing to put more bikes in stations that will likely be emptied.

The first contribution of this paper is a predictive model of availability in BSS based on queuing theory. We model a BSS station as a queue which is subject to two stochastic events: an arrival process for the bikes that are returned, and a departure process for the bikes that are checked out by users wishing to start a journey. Thus each station is modeled in isolation, but the effects from the overall BSS are incorporated into these aggregate processes. For instance, the arrival process subsumes all journeys started from any BSS station, while the departure process accounts for pick-ups by users who had not found an available bike elsewhere.

Modeling a station in isolation simplifies the analysis and the parameter fitting. This approximation is justified by the-

CIKM'15, October 19–23, 2015, Melbourne, Australia.

ACM 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806569.

oretical results stating that in the asymptotic regime of the number of stations going to infinity, under mild assumptions the stations can be viewed as independent of each other [7, 8]. Furthermore, this approximation is also used in more empirical related works which use Bayesian classifiers [9] and time-series analysis [13]. Differently from these, our model is *explanatory* in that the dynamics is explicitly given. Using real data from the *Vélib'* system of the City of Paris, we actually validate that the real dynamics can be accurately captured by time-inhomogenous Poisson arrival and departure processes. To the best of our knowledge, such a validation of a theoretical model of BSS has not been done before. In contrast, Kaltenbrunner *et al.* adopt a deterministic time series [13], while in [9] stochasticity is encoded as a hidden node in the Bayesian network through a Gaussian variable learned from the data. Our model also improves on two other aspects. First, being time-inhomogeneous, it naturally accounts for non-stationary behavior (a limitation of the auto-regressive moving-average model of [13] that was pointed out in [15]). Second, it can predict the whole probability distribution of the number of bikes at a BSS station, instead of a coarser Bayesian classification into occupation ranges as in [9]. Models that take into account the state of neighbouring stations have also been proposed. A time-series model is presented in [15], which encodes stochasticity through an exogenous normally distributed noise term, to be fitted from the data. However, the empirical results do not show significant improvements over neighborhood-unaware predictive models. Guenther and Bradley [12] consider both a time-inhomogeneous population continuous-time Markov chain model and a model that uses regression analysis and errors fitted to a time series. The population model has the same probabilistic behaviour as the queuing model considered in this paper. They use a fluid approximation approach to obtain point estimators for bike availability in each station of the system, whereas we focus on the entire probability distribution for a single station. Furthermore, we validate the core assumptions of the approach, *i.e.*, Poisson arrivals.

The ability to make probabilistic forecasts can improve the usability of the models by both users and providers. A group of users wishing to ride together may obtain an estimate for the probability of finding no fewer bikes than the group size at the origin station [15]. BSS providers, on the other hand, may rank critical stations according to the probabilities of them being empty (or full) at a given time. Unfortunately, all the aforementioned approaches [9, 13, 15, 12] only focus on point estimators.

Acknowledging the stochastic nature of a BSS also has an important implication on the methodology used to measure and compare the accuracy of its predictors. In the bike-sharing literature, the most important accuracy index used so far has been the well-known root mean square error (RMSE) between the point predictor and the actual observed outcome. This is also the metric used in an ongoing initiative aimed at comparing predictive models of BSS [1]. Our second main contribution, of a methodological nature, is to challenge the use of this metric by putting forward two arguments. First, we claim that it does not properly account for the inherent stochasticity in the system. In the ideal case, an RMSE of zero would indicate a perfect predictor; larger values can be used to rank the accuracy of different predictors. Instead in this paper we show an analysis concluding that, when applied to a BSS, even a point predictor with perfect information about the dynamical evolution of the system *cannot* yield zero RMSE. Second, we observe that the point predictors available so far may not always be informative for practical applications such as the deployment of a recommendation system. In this case, for instance, the user is not interested in the actual number of bikes available at some point in the future; rather, she is interested in being able to pick up a bike with high probability. This information cannot be recovered directly from point estimators.

To overcome these difficulties, we propose new metrics for BSS based on scoring rules, known to be appropriate to evaluate the accuracy of probabilistic predictors and to relatively compare them [11]. We use classical scoring rules to evaluate our ability to predict the number of available bikes. We propose a new scoring rule to evaluate our ability to recommend the feasibility of a journey based on a tuneable threshold on the probability of finding at least one bike at the origin station. We apply this new methodology to evaluate the accuracy of our queuing model using a real dataset from *Vélib'*. We show that for prediction horizons up to 5 hours, it significantly outperforms other predictors such as the Last-Value predictor (where the prediction is equal to the last observed value) and the Historic-Trend predictor (where the prediction equals the historical averages).

Contributions. To summarize, this paper makes the following contributions.

- We discuss an explanatory time-inhomogeneous queuing theoretic model of a BSS station and successfully validate it through fitting of the arrival and departure processes to Poisson distributions using real data from the *Vélib'* BSS of the City of Paris.
- We show that RMSE is not an appropriate metric for journey planning. Instead, we propose new metrics based on scoring rules which can capture provider-related measures (e.g., estimating the number of bikes at a station) as well as user-oriented ones (e.g., estimating the probability that the origin station of a journey is not empty).
- We evaluate our predictor on data from *Vélib'*, showing that it outperforms other point predictors for prediction horizons of 2 to 5 hours, thus advocating its use as a probabilistic recommendation system for BSS.

Paper structure. We discuss the queuing model in Section 2, and validate the model against historic data in Section 3. We show that RMSE in not relevant for our metric in Section 4. We introduce a new metric based on scoring rules in Section 5. We conclude in Section 6.

2. MATHEMATICAL MODEL AND DATASET

2.1 Predicting Trips Feasibility

We focus on a BSS that consists of N stations located in different areas of a city. Each station $i = 1, \ldots, N$ has a fixed *capacity* κ_i, which corresponds to the maximum number of bikes that can be parked at this station. The number of bikes that are available for rental at station i at time t is denoted by $X_i(t) \in \{0, \ldots \kappa_i\}$ and the occupancy of all stations

at t is represented by a vector $\mathbf{X}(t) = (X_1(t), \ldots, X_N(t))$. Two types of events modify the state $X_i(t)$: If $X_i(t) > 0$, a bike can be picked up at station i (this modifies $X_i(t)$ into $X_i(t) - 1$); and if $X_i(t) < \kappa_i$, a bike can be returned at the station (therefore increasing $X_i(t)$ by one). It should be clear that $\mathbf{X}$ is a complex stochastic process: these events are influenced not only by external factors like weather and traffic conditions, but also by endogenous variables such as the occupancy of other stations in the network.

Now assume that at time t, a user decides that she wants to start a trip at time $t+h$ from station i to go to station j. If she had a perfect knowledge of the occupancy of the stations at any time, her decision to use the system would only depend on whether there are enough resources available to complete the trip successfully. That is, she is interested in knowing if station i is non-empty at the time of departure and station j is non-full at the time of arrival. If τ_{ij} is the travel time from i to j, the trip is *successful* if the conditions $X_i(t+h) > 0$ and $X_j(t+h+\tau_{ij}) < \kappa_j$ are satisfied. In practice, the user has only access to the information up to time t, denoted by $\mathcal{F}_t$. Given the stochastic nature of the evolution of $\mathbf{X}$, the problem from the user's perspective therefore reduces to estimate the following probability:

$$\mathbb{P}\left(X_i(t+h) > 0 \wedge X_j(t+h+\tau_{ij}) < \kappa_j \mid \mathcal{F}_t\right). \qquad (1)$$

In what follows, we aim to define a method to provide the user with a reliable prediction for the feasibility of her trip. We rely on a stochastic model introduced in the next section.

2.2 A Queuing Model for a Single Station

Throughout this paper, we focus on the behavior of one station is isolation. We consider the following Markovian model for a station's behavior: Users arrive at station i to pick up bikes according to a time-inhomogeneous Poisson process[1] of intensity $\mu_i(t)$, and bikes are returned to station i according to a time-inhomogeneous Poisson process of intensity $\lambda_i(t)$. Using Kendal's notation for queuing networks [2], one station is modeled as a time-inhomogeneous $M/M/1/\kappa_i$ queue, represented in Figure 1.

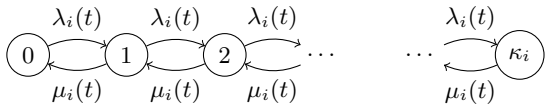


Figure 1: Time-inhomogeneous Markovian model of a station i.

This model makes two approximations. The first one is that the transitions of our system are memory-less and that the arrival processes of users and bikes can be represented as a Poisson process. While this might seem restrictive, we will see in Section 3.2 that this behavior is close to what happens in reality. Moreover, Markovian models have been successfully used to model bike-sharing systems, for example in [7, 10, 12]. The second assumption is that the state of a particular station does not depend on the state of the other ones. Again, this is not true in practice since when a station is empty, no bike can depart from it, therefore reducing the arrival rate at other stations. An alternative, more realistic assumption, would be to consider a process for each trip from an origin i to a destination j, with departure and arrival intensities $\lambda_{ij}(t)$ and $\mu_{ij}(t)$ for each process. However this greatly complicates the model and the parameter fitting for little gain. Moreover, the great number of stations usually involved in BSS such as *Vélib'* allows us to apply the asymptotic analysis techniques introduced in [8], which argue that the isolation approximation becomes exact as the number of stations goes to infinity. Therefore in this paper, we restrict ourselves to a model in which each station behaves as an independent $M/M/1/\kappa$ queue.

Under this assumption, the probability (1) simplifies into the independent computation of $\mathbb{P}\left(X_i(t+h) > 0 | X_i(t)\right)$ and $\mathbb{P}\left(X_j(t+h+\tau_{ij}) < \kappa_j | X_j(t)\right)$. The problem is symmetric between the departure (number of bikes available) and the arrival (number of free parking stands) stations. Hence, in the remainder of this paper, we focus on predicting the non-empty probability.

2.3 Deterministic and Probabilistic Predictors

In this section, we describe the predictors that we will compare later. A probabilistic predictor issued at a time $t \in \mathcal{T}$ for a time $t+h \geq t$ is a guess about the probability distribution of $\mathbb{P}\left(X_i(t+h) = y \mid X_i(t) = x\right)$, and can be written as $p_i(y|x,t,h)$. Here, $y \in \{0, \ldots, \kappa_i\}$, $p_i(y|x,t,h) \in [0,1]\ \forall y$ and $\sum_y p_i(y|x,t,h) = 1$. In words, $p_i(y|x,t,h)$ is the estimated probability that there are y bicycles at station i at time $t+h$, knowing that there are x bicycles at time t.

We will now discuss some commonly chosen predictors.

Last-value (LVP). The last-value predictor assigns all probability mass to the number of bicycles in the last observation:

$$p_i(y|x,t,h) = \begin{cases} 1 & \text{if } y = x, \\ 0 & \text{otherwise.} \end{cases}$$

This predictor depends on the time t at which the prediction is issued but not on the prediction horizon h.

The "historical" predictor (HP). The historical predictor uses statistical information about the past. We decompose each day into intervals of 15min. Our dataset is composed of measurements $x_i(d,s)$, where i is a station, $d \in D$ is one day in the set of days for which we have measurements and $s \in S$ represents a time interval. Let $|D|$ denote the number of elements in D. The historical predictor is then given by

$$p_i(y|x,t,h) = \frac{1}{|D|}\#\{x_i(d, \text{time of day}(t+h)) = y\}.$$

This predictor does not depends on t directly. It only depends on which time-interval of the day $t+h$ belongs to. The current state x of the station does not influence this predictor.

The "queuing model" predictor (QMP). As we discussed in Section 2.2, we model the behavior of each station i as an $\mathrm{M/M/1/}\kappa_i$ queue with varying arrival or departure rate. If we know these two parameters, we can compute the probability that a station is full or empty at a given time in the future. Indeed, the transition kernel of the Markov chain for

[1] A time-inhomogeneous Poisson process of intensity $\lambda(t)$ is defined by two properties. First, for any $t_1 < t_2$, the number of events A_{t_1,t_2} a time interval $[t_1, t_2]$, follows a Poisson distribution of parameters $\int_{t_1}^{t_2} \lambda(s)ds$. Second, when $t_1 < t_2 < t_3 < t_4$, then A_{t_1,t_2} and A_{t_3,t_4} are independent

the station i is $Q(\lambda_i(t), \mu_i(t))$, where $Q(\lambda, \mu)$ is the following matrix:

$$Q(\lambda,\mu) = \begin{pmatrix} -\mu & \mu & & & \\ \lambda & -(\lambda+\mu) & \mu & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda & -(\lambda+\mu) & \mu \\ & & & \lambda & -\lambda \end{pmatrix}. \quad (2)$$

Note that to simplify the notation, we often denote by $Q_i(t)$ the quantity $Q_i(t) = Q(\lambda_i(t), \mu_i(t))$.

By using the Kolmogorov equation, the probability that the station has y bikes at time $t+h$ given that this station has x bikes at time t is:

$$p_i(y|x,t,h) = \exp\left(\int_0^h Q(t+s)ds\right)_{x,y},$$

where $\exp(M)$ is the matrix exponential of M and $\exp(M)_{x,y}$ is its term at row x and column y. In Section 3.1, we discuss how to estimate $\lambda_i(t)$ and $\mu_i(t)$ using the historical dataset.

"Always go" predictor. The "always go" predictor predicts that the station will be non-empty. That is, it returns a probability distribution for the occupancy of the station such that for all x, t, h, $p_i(0|x,t,h) = 0$. It disregards any information about the system and is independent of t and h. It models a situation in which the user plans to use the system but has no *a priori* knowledge about the system.

2.4 Dataset

In all numerical evaluation, we use historical data from the *Vélib'* system in Paris collected between 1 October 2013 and 31 December 2014 provided by the operator of *Vélib'*. Our dataset contains two types of data, one describing the availability at all the stations of the system and the other containing the individual trip information. The station measurements are made every minute and specify the number of available bikes and free slots in each station. In the other dataset each bike move is recorded with an indication of the reason for the move (bike rental by a user, maintenance or reallocation), the stations of origin and destination as well as the departure and arrival dates and times. We use this information to compute a historical predictor and estimate the (bike) departure and arrival rates, for each minute of a typical working day or weekend day (see the details in the next section).

In the numerical evaluations, we issue a new forecast each day of the week at $t \in \{7\text{am}, 11\text{am}, 3\text{pm}, 6\text{pm}\}$ for prediction horizons from 5min to 10 hours. To simplify notation, in Sections 4 and 5, we write that we issue forecast at a time $t \in \mathcal{T}$. In fact, each element of the set $\mathcal{T}$ corresponds to a prediction issued for one particular station, at one of the times (7am, 11am, 3pm or 6pm) of a particular day. What we call the average performance of a predictor is the average over these $|\mathcal{T}| \approx 2 \cdot 10^6$ values.

3. MODEL VALIDATION

3.1 Parameter Estimation

We estimate the intensities $\lambda_i(t)$ and $\mu_i(t)$ defined in Section 2 by using the dataset. As written in Section 2.3, each measurement can be written as $x_i(d,s)$, where $d \in D$ is a day and $s \in \mathbb{N}$ represents the time of day. The set D of days

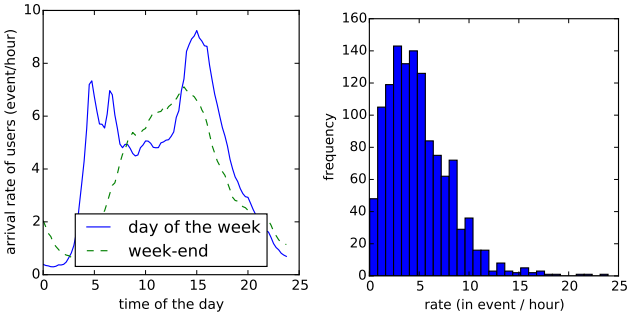


Figure 2: Station use intensities: arrival of users as a function of the time of the day (left) and histogram of arrivals between 10am and 12am (right).

is such that the intensities are assumed to be the same (e.g., all weekdays and roughly in the same season).

So far, we have not imposed any requirements on $\lambda_i(t)$ and $\mu_i(t)$. We consider that intensities are *piecewise constant* (also called *step functions*) and depend only on the time of day and whether this day is during the weekend or not. For one day, the function $\lambda_i : [0, 24h] \to \mathbb{R}$ is piecewise constant with n steps/pieces on $\mathbb{R}$ if there exist times $t_1, \ldots, t_{n-1}$ and step heights $l_{i1}, \ldots, l_{in}$ such that

$$\lambda_i(x) = l_{ij} \quad \text{if} \quad t_{j-1} < x < t_j,$$

with $t_0 = 0$ and $t_n = 24h$, and t_j, $l_{ij} \in \mathbb{R}$ $\forall j$. Similarly, μ_i is piecewise constant if

$$\mu_i(x) = m_{ij} \quad \text{if} \quad t_{j-1} < x < t_j,$$

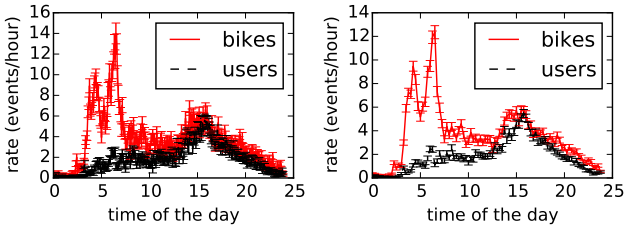
with $m_{ij} \in \mathbb{R}$ $\forall j$. We consider that the times $t_1, \ldots, t_{n-1}$ are evenly distributed (all intervals have the same duration) and we estimate the step heights from the dataset. A naive estimator for l_{ij} (in seconds) based on the $|D|$ days observed would be to count the number of observed arrivals in the $|D|$ corresponding intervals and divide this by $(t_j - t_{j-1})|D|$. This, however, ignores the fact that we do not observe arrivals that take place when the station is full. We say that a second during the observation $x_i(s,d)$ is 'valid' for the purpose of estimating arrival intensities if station i is non-full; the resulting estimator for the arrival rate is

$$\hat{l}_{ij} = \frac{\sum_{d \in D} \#\{\text{arrivals on } [t_{j-1}, t_j] \text{ on day } d\}}{\sum_{d \in D} \#\{\text{'valid' seconds on } [t_{j-1}, t_j] \text{ on day } d\}}. \quad (3)$$

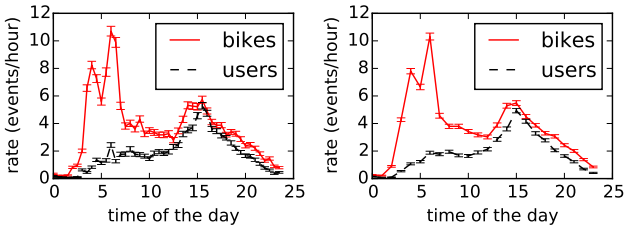
A similar argument holds for $\hat{m}_j$, departures and seconds during which the station is non-empty.

The estimated intensities l_{ij} and m_{ij} vary to a considerable degree among the 1225 stations in the dataset. For example, the variation in terms of arrival intensities is displayed in the right-hand figure of Figure 2. Still, some meaningful trends can be observed, especially if the measurements are partitioned into a 'weekday' and a 'weekend' part. This can be seen in the left-hand figure of Figure 2. In particular, the weekdays have two rush-hour peaks whereas the weekends have a single peak in the afternoon.

Figure 3 displays how the choice of the number of steps impacts the estimated intensities. When there are many steps (*e.g.*, $t_i - t_{i-1} = 5\text{min}$ as in the top-left figure), the estimates have considerable variance: given a dataset of a fixed size, more windows mean fewer observations per win-

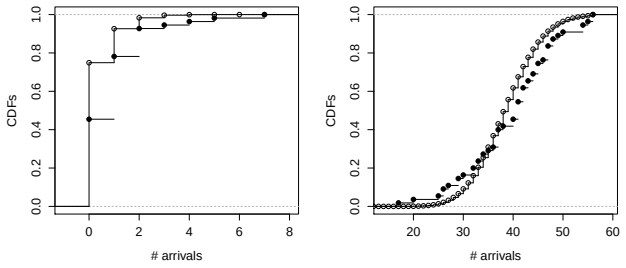


(a) Steps of 5min (b) Steps of 15min



(c) Steps of 30min (d) Steps of 60min

Figure 3: Estimated intensities using different time granularities for the "Pecqueur" station. The error bars show the 95% confidence interval of the estimated parameter of the Poisson distribution.



(a) Station "Pecqueur" (b) Station "Gare du Nord"

Figure 4: Comparisons between the theoretical (white dots) and empirical (black dots) CDFs for bike arrivals at two stations, in time window 6. The K-S test statistic is the maximum (vertical) distance between a pair of white and black dots in the graphs. The station on the left has a higher K-S test statistic ($K = 0.3815$) than the one on the right ($K = 0.1633$). There were 53 total arrivals at Pecqueur in this time window throughout the observation period, and 2186 at Gare du Nord (2187 if we account for a brief period of station fullness).

dow, which in turn means more variance. However, if the windows are very large (as in the bottom-right figure, where there are only 24 windows per day), too much information may be lost. For example, the fact that the morning rush hour seems to have two peaks is less clear in the bottom-right panel of Figure 3. In the numerical evaluation, we divided a day in 96 intervals of 15min and we focus only on week-days.

3.2 Model Validation

Alongside parameter estimation, we use the data to validate the assumptions underlying the queuing model. The central assumption underlying the choice to model each station as a time-inhomogeneous $M/M/1/\kappa$ queue is that the bicycle arrivals and departures form Poisson processes. There are several ways to validate this. For example, one could verify whether inter-event times are exponentially distributed and uncorrelated. In this paper, we check whether on each time interval on which the arrival and departure rates are constant, the distribution of arrivals/departure follows a Poisson distribution. Let

$$A_j = \#\{\text{arrivals during } [t_{j-1}, t_j]\}, \quad x, y \in \mathbb{R},$$

and let D_j be similarly defined for the departure process. Then if $\lambda(t) = \lambda \ \forall t \in [t_{j-1}, t_j]$ for some $\lambda \in (0, \infty)$, it should hold that

$$\begin{aligned} F_{A_j}(x) &= \mathbb{P}(A_j \leq x) \\ &= \sum_{z=0}^{x} e^{-\lambda(t_{j-1}-t_j)} \frac{(\lambda(t_{j-1}-t_j))^z}{z!}. \end{aligned} \tag{4}$$

Again, this is analogous for D_j and a constant μ. If we partition the entire time period of interest into n pieces as discussed in Section 3.1, then we would obtain for all j one measurement a_j of A_j per day. Again, this assumption is invalid if there are parts of the day where the bicycle station is full (for arrivals) or empty (for departures). There are two remedies: we can discard days when the station becomes full/empty, or we can assume that if the station is non-full/non-empty some fraction p of the time, we rescale the measurements by p (so if we observe x arrivals in window j on some day and the station is full half the time, we would record $2x$ measurements). The latter choice will have an impact on the confidence levels of related statistical tests; still, this is what we will use in this paper.

Having obtained measurements $(a_{j,d})_{d \in D}$ of the arrivals during time window j on day d, the question is whether the resulting empirical CDF given by

$$\hat{F}_{A_j}(m) = \frac{1}{\#D} \#\{a_{j,d} \leq m\}$$

is 'sufficiently' close to $F_{A_j}(m)$ as defined in (4). Several measures of the distance between F_{A_j} and $\hat{F}_{A_j}$ — called 'goodness-of-fit' measures — exist, including the Cramér-von Mises criterion and the Kolmogorov-Smirnov (K-S) test statistic. We focus on the latter, given by

$$K = \max\left\{|\hat{F}_{A_j}(m) - F_{A_j}(m)|, m \in \mathbb{N}\right\}.$$

A statistical test for the null hypothesis that $F_{A_j}(m)$ is the distribution underlying $\hat{F}_{A_j}(m)$ can be performed based on K. This test is implemented for discrete distributions in the `dgof` package of the statistical package R. Note that we do not actually conduct the K-S test in this paper; we just use its criterion as an indicator for prediction accuracy.

The K-S test statistic is illustrated in Figure 4. For two stations, both the empirical and Poisson distribution functions have been displayed. The K-S test statistic is the maximum difference between the two functions. As can be seen, the K-S test statistic has a lower value (*i.e.*, a better fit) for the station on the right than for the station on the left.

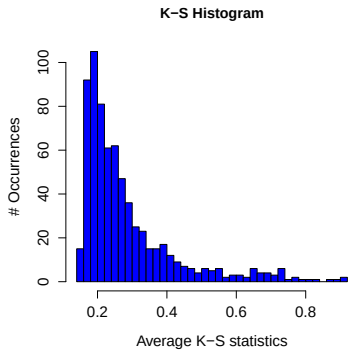


Figure 5: Histogram of the values of the average K-S statistics for the bike arrivals during the one-hour time windows between 5AM and 8PM.

In general, stations for which we have many observations have a better fit with the Poisson distribution. Computation of K-S test statistics, or even an average of K-S statistics over all (or just the daytime) time windows for a station gives an indication of how well the prediction algorithm is going to perform. An example of this is Figure 5, which is a histogram of the average K-S statistics for the bike arrival processes during the one-hour time windows between 5AM and 7PM. The stations for which our approach can be expected to do best will be those stations corresponding to the peak on the left. To mitigate seasonal effects, we only considered observations between 1 December 2013 and 28 February 2014. Stations for which there was at least one one-hour window without arrivals between 5AM and 8PM throughout the entire observation period, or for which the station was full or empty for a whole hour during this period, were discarded, leaving 692 stations.

The median K-S statistic in Figure 5 is 0.23579 (see Figure 4 for a visual interpretation of the K-S statistics). Bearing in mind that even small weather effects would violate the assumption of Poisson distributions, we can conclude that the model assumptions are valid to a reasonable degree.

4. A CRITIQUE OF DETERMINISTIC FORECASTS AND RMSE

Most of the work dealing with prediction for bike-sharing systems, *e.g.*, [9, 12, 13, 15], focuses on providing a single value for $X_i(t)$, whose performance is generally evaluated using the root-mean-square-error (RMSE). The recently initiated bike-sharing prediction challenge [1], also uses this metric: Participants are expected to submit a deterministic time series that represents the estimated number of bikes at a given station in the future. In this section, we show that regardless of their quality, the RMSE of such predictors will always be large due to the stochastic nature of bike-sharing systems. For instance, we present evidence that the best RMSE one can expect from a single-valued estimator for a prediction 2 hours in advance will lie between 10% and 20% of the capacity of a station.

4.1 Deterministic Prediction and RMSE

Assume we want to determine the occupancy of the stations h hours in advance, using all the information available up to now. A deterministic prediction of the vector $\mathbf{X}$ issued at t for the time $t+h$ is a vector $\mathbf{Y}^t(t+h)$, where for all $i = 1, \ldots, N$, $\mathbf{Y}^t_i(t+h) \in [0; \kappa_i]$ is a single real-valued number that corresponds to the predicted number of bikes at time $t+h$ for the station i.

A classical way to evaluate the performance of a deterministic prediction [1, 9, 12, 13, 15] is to compute the root-mean-square-error (RMSE) of a predictor, defined as

$$RMSE(\mathbf{Y}, h) = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \|\mathbf{X}(t+h) - \mathbf{Y}^t(t+h)\|_2^2},$$

where $\|\mathbf{X} - \mathbf{Y}\|_2^2 = \sum_i (X_i - Y_i)^2$ and $\mathcal{T}$ is the set of time steps at which predictions are made. Note that the RMSE depends on the predictor used ($\mathbf{Y}$) and on the time horizon of the prediction, h.

4.2 A Lower Bound on the RMSE

The smaller the RMSE, the closer our prediction is to the true value. It is tempting to say that the best predictor should have an RMSE of 0. However, we have seen in Section 2 that bike-sharing systems exhibit a stochastic behavior. We now show that this stochasticity implies that the best predictor will have an RMSE of 1 bike at 5min and 3 bikes at 1h for typical parameter values, even if we have perfect knowledge of the parameters of our model (see Figure 6a). Note that these values are consistent with the ones found in [15, Table I] in which the authors show that all their predictors have this performance. Note that we also obtain similar figures for the Paris dataset (see Figure 6b).

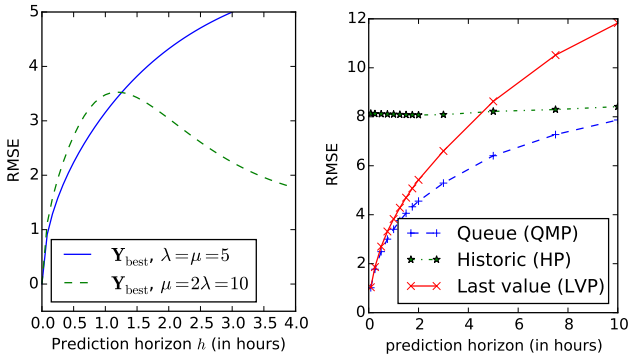
We consider a simple scenario with one station for which $\lambda(t) = \lambda$ and $\mu(t) = \mu$ are constant in time. The quantity $X(t)$ behaves as a time-homogeneous continuous time Markov chain of transition kernel $Q = Q(\lambda, \mu)$, as defined in Equation (2). In particular, we have

$$\mathbb{P}\left(X(t+h) = j \mid X(t) = i\right) = \exp(Qh)_{ij}$$

In this case, the deterministic predictor $\mathbf{Y}^t_{\text{best}}(h)$ that minimizes the RMSE is $\mathbf{Y}^t_{\text{best}}(h) = \sum_{i=0}^{\kappa} (\exp(Qh)_{X(t),i})i$.

In Figure 6a, we plot the RMSE of the best predictor. We consider two scenarios, in which the capacity of the station is $\kappa = 20$ bikes and such that at the time $t \in \mathcal{T}$ when the predictions are made, $X(t) = 10$ bikes. In the first scenario, the flow at the station is balanced: $\lambda = \mu = 5$ bikes/hour while in the second scenario: $\lambda = 5$ bikes/hour and $\mu = 2\lambda$.

We draw two observations from these figures. First, for both scenarios, the RMSE are around 0.9 or 1.1 for a prediction at horizon $h = 5$min and around 3.5 for a prediction at horizon $h = 1$h. This implies that even someone who has perfect knowledge of the parameters λ and μ of the system cannot predict the number of bikes in $1h$ with a RMSE less than 3.5 bikes. In particular, this shows that the predictors presented in [15] all produce values that are close to the best that can be achieved. The second remark is that the RMSE is not sufficient to provide the users with a precise indication on whether a trip is feasible. In particular, a user who wants to depart from a station is interested in knowing if there will be 0 or at least 1 bike but has little interest in the real value. These numbers are hard to guess from deterministic predictors. For example, in the unbalanced scenario, the predictions at $1h$ are $\mathbf{Y}_{\text{best}}(2h) = 2.50$ which suggest that the trip would be possible. Another tentative approach would be to use a linear predictor $\mathbf{Y}^t_{\text{linear}}(h) = ((X(t) + (\lambda - \mu)h) \vee 0) \wedge \kappa$,



(a) Artificial model (b) Paris dataset

Figure 6: RMSE for the best-predictors on an artificial model (left) and on the Paris dataset (right).

which corresponds to a linear evolution of the number of bikes truncated at 0 or κ when the capacity is exceeded. This would give $\mathbf{Y}_{\text{linear}}(2h) = 0$ and suggest that the trip is not feasible. A true calculation shows that the probability of empty is 0.34, but neither the two predictors proposed above is able to provide this crucial information.

4.3 RMSE on the *Vélib'* data

In Figure 6b, we compare the RMSE of the deterministic predictors based on the queuing model introduced in Section 2.2, as well as the ones based on historical data and the last-value. The deterministic queuing model predictor is $Y^t(t+h) = \sum_{i=0}^{\kappa}(\exp(Qh)_{X(t),i})i$ and the historic is the average number of bikes observed is the past at hour $t+h$.

For all prediction horizons h, our queuing model provides the best RMSE and converges to the RMSE of the historic model as h increases. Last-value outperforms the predictor from historic data up to $h = 5$ hours but it deteriorates significantly as the horizon grows, reaching values worse by up to 50% for a large prediction horizon. In fact, we will see in Section 5 that when one wants to assess the feasibility of a trip, last-value is beaten by historic as soon as the prediction horizon exceeds $h = 1$ hour. This again reflects the fact that the RMSE is not an appropriate metric when the objective is to determine whether a trip will be successful: LVP has a much smaller RMSE than HP for a time horizon of 2h but is worse at predicting successful trips (Figure 9) because in the latter case, knowing if there will be exactly 10 or 15 bikes in a station is less important whereas knowing if there will be 0 or 1 is critical. As we will see in the next section, we need a different metric to easily express these estimations and measure probabilistic forecasts.

5. PROBABILISTIC FORECASTS AND SCORING RULES

Bike-sharing systems exhibit stochastic behavior. Hence, an alternative to using a deterministic predictor is to forecast the probability distribution of the number of available bikes at the station in the future. In this section, we first show how to evaluate probabilistic predictors by using proper scoring rules. We then develop two scoring rules to assess the performance of QMP in terms of its probabilistic predictions.

5.1 Proper Scoring Rules

A probabilistic forecast for a value X can be represented by a probability distribution $\mathbf{P}$. A probabilistic forecast contains information about the value that we think X will take but also a measure of confidence in this forecast. A natural way to evaluate a forecast is to use proper scoring rules. We now recall briefly the mathematical definition of a proper scoring rule; see [11] for a more complete exposition.

Assume that we want to forecast a random variable X that takes values in a set $\mathcal{X}$. A scoring rule is a function $S(\mathbf{P}, .) : \mathcal{X} \to \overline{\mathbb{R}}$ that associates to each $i \in \mathcal{X}$ a score $S(\mathbf{P}, i) \in \mathbb{R} \cup \{-\infty\}$. The score of a forecast $\mathbf{P}$ is defined by its average score $\mathbb{E}(S(\mathbf{P}, X))$. A scoring rule is said to be proper if, when X is generated according to a probability distribution $\mathbf{Q}$, the score is maximized by the distribution $\mathbf{Q}$. In other word, we have for any distribution $\mathbf{P}$ on $\mathcal{X}$:

$$\mathbb{E}(S(\mathbf{Q}, X)) \geq \mathbb{E}(S(\mathbf{P}, X)). \tag{5}$$

It is strictly proper if $\mathbf{Q}$ is the only maximizer of (5). Note that scoring rules are defined without a priori knowledge on the distribution $\mathbf{Q}$: it is not necessary to know $\mathbf{Q}$ to define a proper scoring rule and to compute the score $S(\mathbf{P}, X)$. This is important as in practice the distribution $\mathbf{Q}$ is unknown (it is the one we try to forecast).

A proper scoring rule encourages honesty: a forecaster knowing the true distribution $\mathbf{Q}$ of $\mathcal{X}$ should issue a forecast $\mathbf{P} = \mathbf{Q}$. If one is very confident that $X \approx x$, then the issued predictor $\mathbf{P}$ will be concentrated on a single value x. In that case, the score $S(\mathbf{P}, X)$ will be high when $X \approx x$ but very low when X is far from x. On the other hand, when one has a low confidence in the prediction, $\mathbf{P}$ will give weight to many values and $S(\mathbf{P}, X)$ will be neither high nor low regardless of the actual value of X.

Skill score. When evaluated over a trace of measurement, the performance of a probabilistic forecast is measured in terms of average score over all forecasts and realizations:

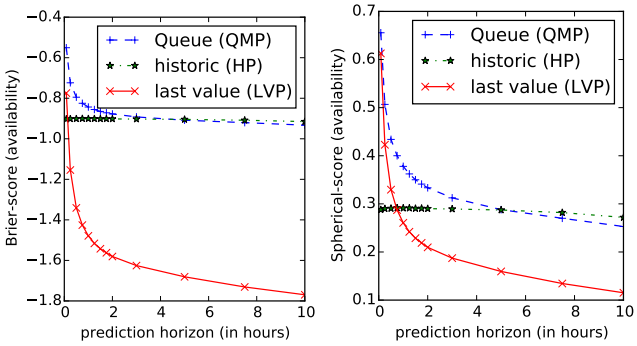
$$\text{Score}(\mathbf{P}, h) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} S(\mathbf{P}^t(t+h), \mathbf{X}(t+h)). \tag{6}$$

The higher the score, the better the forecast.

5.2 Scoring for the Number of Available Bikes

The predictors described in Section 2.3 provide a deterministic or probabilistic estimation of the number of bikes in a station in the future. Each predictor is a probability distribution over a finite set $\{0, 1 \ldots, \kappa_i\}$, where κ_i is the capacity of the station i considered. Note that a deterministic predictor is a degenerated probabilistic predictor that assigns all the probability mass to a single value. We evaluate numerically these predictors on the Paris dataset by using classical scoring rules. The results are reported in Figure 7.

The most classical scoring rules are Brier score, the spherical score and the logarithmic score. Given a forecast $\mathbf{P}$ and a realization i, Brier score assigns a score $S(\mathbf{P}, i) = 2p_i - \sum_j p_j^2 - 1$, the spherical score assigns $S(\mathbf{P}, i) = p_i/(\sum_j p_j^2)^{1/2}$ and the logarithmic score assigns $S(\mathbf{P}, i) = -\log p_i$ to $\mathbf{P}$, respectively. In Figure 7, we report the average score of the three predictors defined in Section 2.3 as a function of the prediction horizon for Brier and the spherical score. We choose not to report the performance of the logarithmic score for two reasons: First, because the performance of the QMP



(a) Brier score (b) Spherical score

Figure 7: Average Brier score and spherical score when predicting the number of bikes available (the higher the score, the better is the predictor).

and of HP exhibit the same trend, and second because the average score of LVP is always $-\infty$ with this score (LVP is a deterministic predictor, so when the actual number of bikes is not exactly equal to the prediction, the score is $-\infty$).

For both scoring rules, the shape of the curves are similar: the QMP has the best score for a prediction up to $h = 5$ hours in advance. After 5 hours, HP becomes slightly better. This seems to indicate that knowing the current number of bikes in a station is only important for prediction up to 2 to 3 hours in advance. Moreover, for both scoring rules, LVP quickly has the worst score (for prediction horizons larger than 10min (Brier) or 45min (spherical)). This comes from the deterministic nature of LVP, leading to an over-confidence in wrong predictions that tend to be heavily penalized by these scoring rules.

5.3 Scoring for the Availability of One Bike

Predicting the number of available bikes is important from the system operator's point of view, for example when she decides how to redistribute bikes in the system. However, this metric is less useful from a user's point of view: A user who wants to take a bike from a station is only interested in knowing if there is at least 1 bike available but is less interested in knowing, *e.g.*, if there are 5 or 15 bikes. In this section, we design a scoring rule to evaluate a user-centered prediction, which aims at predicting if there will be "no bikes" or if there will be "1 or more bikes" available at a given station at a given time in the future.

To construct our scoring rule, we assume that a user has preferences that can be modeled by a utility function. If a user decides to go to a station and finds a bike available, she gets a utility $U(\text{Go}, \text{OK})$ greater than $U(\text{Go}, \emptyset)$, the score she would get if she goes and the station is empty. If she decides not to go, her utility is $U(\text{No go}, \emptyset) \geq U(\text{No go}, \text{OK})$, indicating that she can be frustrated by knowing that in fact her trip was possible. If the user knows the probability p for a station to be non-empty, she will decide to go to the station to see if there is a bike if $p \geq p^*$, where p^* is equal to:

$$p^* = \frac{S(\text{Go}, \emptyset) - S(\text{No go}, \emptyset)}{S(\text{Go}, \emptyset) + S(\text{No go}, \text{OK}) - S(\text{Go}, \text{OK}) - S(\text{No go}, \emptyset)}.$$

Theorem 2 of and Section 3.2 of [11] show that for any utility U such that $U(\text{Go}, \text{OK}) \geq U(\text{Go}, \emptyset)$ and $U(\text{No go}, \emptyset) \geq U(\text{No go}, \text{OK})$ the following score is a proper scoring rule:

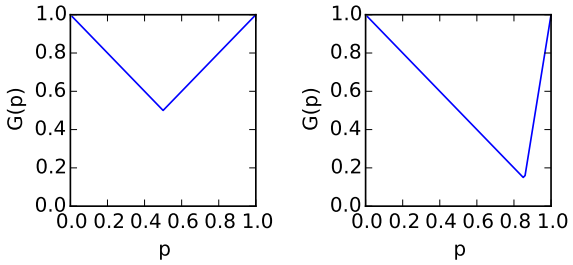
$$S(p, i) = \begin{cases} U(\text{Go}, i) & \text{if } p \geq p^* \\ U(\text{No go}, i) & \text{if } p < p^* \end{cases} \quad (7)$$

More generally, if $G : [0, 1] \to \overline{\mathbb{R}}$ is a convex function, then a scoring rule that assigns the following scores is proper:

$$S(p, \text{OK}) = G(p) + (1 - p)G'(p) \quad (8)$$

$$S(p, \emptyset) = G(p) - pG'(p). \quad (9)$$

Conversely, any proper scoring rules have the form (8)-(9). The function G corresponding to the score (7) is represented in Figure 8 for the parameters $U(\text{Go}, \text{OK}) = U(\text{No go}, \emptyset) = 1$, $U(\text{No go}, \text{OK}) = 0$ and two possible values of $U(\text{Go}, \emptyset)$.



(a) $U(\text{Go}, \emptyset)$=0 (p^*=0.5) (b) $U(\text{Go}, \emptyset)$=−5 (p^*≈0.86)

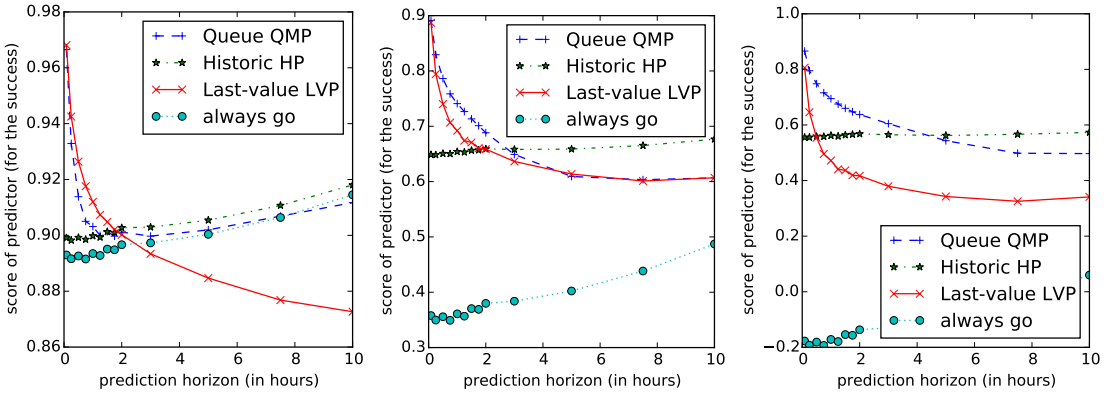
Figure 8: Example of functions G of Equation (8) that correspond to the scores given by Equation (7).

We evaluate numerically the average score on the Paris dataset for the three predictors given in Section 2.3 and the scoring rule (7). We plot the average score of the different predictors as a function of the prediction horizon in Figure 9. We set the parameters $U(\text{Go}, \text{OK}) = U(\text{No go}, \emptyset) = 1$ to normalize the values at 1: a perfect predictor will get a average score of 1. Moreover, we set $U(\text{No go}, \text{OK}) = 0$ and we examine three possible values of $U(\text{Go}, \emptyset) \in \{0, -5, -10\}$. These may indicate increasing degrees of "conservativeness" of a recommendation system, which more heavily penalizes false positives. In practice, we believe that it is reasonable to assume that $U(Go, \emptyset)$ is closer to -10 than 0 to capture that a user is more frustrated by going at a station and not finding a bike than deciding to not go to this station and realizing later that there was in fact a bike available.

First, we remark that in Figure 9, the performance of HP and of the always-go policy depends on the prediction horizon. This comes from the fact the plotted values are averages over predictions issued at 7am, 11am, 3pm and 6pm. Hence, a value at prediction horizon of 10h is the average over values at 5pm, 9pm, 1am and 4am which is different from the average over values at 7am, 11am, 3pm and 6pm. Figure 9a indicates that, on average, a user is more likely to find a bike in the former case.

We note that when $U(\text{Go}, \emptyset) = 0$ (Figure 9a), the score obtained by a predictor corresponds to the proportion of accurate information it gives to the user. Thus, from the score the always-go predictor we can see that, on average, the stations were non-empty 89% of the time.

For a short horizon (5-15min), both LVP and QMP make relatively accurate predictions compared to the other predictors. However the performance of LVP significantly decreases with the prediction horizon, while QMP remains



(a) $U(\text{Go}, \emptyset)$=0 (b) $U(\text{Go}, \emptyset) = -5$ (c) $U(\text{Go}, \emptyset) = -10$

Figure 9: Average score (defined in Equation (7)) for measuring the ability of a predictor to predict the feasibility of a trip as a function of the prediction horizon (higher scores mean better predictors).

competitive. In particular in the case $U(\text{Go}, \emptyset) = 0$, notice how QMP separates from LVP and catches up with the better middle and long-term predictors when the horizon becomes large. In the latter case (*i.e.* more than 3 hours), HP is sufficient to give an accurate prediction to the user. This emphasizes how the main source of information used by a predictor influences its "optimal" length of the prediction horizon: The more heavily it relies on the current state of the system (cf. LVP), the more appropriate it will be for short-term predictions. Instead a predictor based on the average behavior of a station such as HP will exhibit better performance for large prediction horizons, because the future availability becomes independent of the current occupancy.

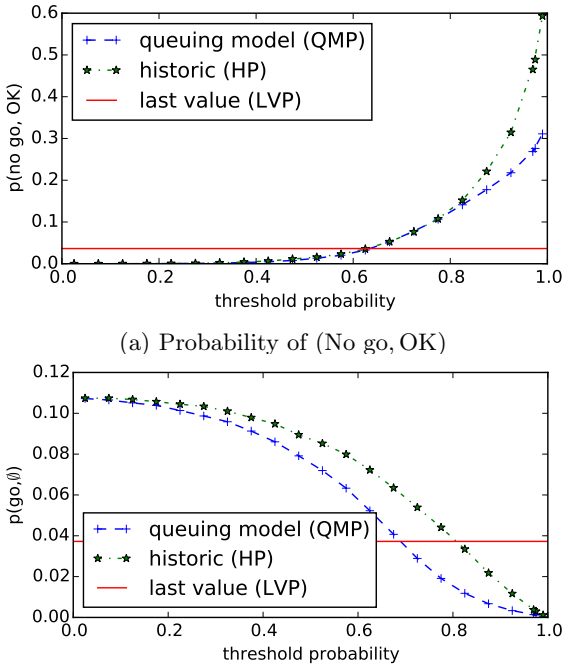
While QMP only slightly outperforms LVP and is comparable to HP in their respective optimal prediction windows, it clearly stands out when it comes to middle-term predictions, *i.e.* for horizons between 15 minutes and 3 hours. In particular, by comparing the accuracy of QMP as a function of $U(\text{Go}, \emptyset)$ indicates that it is better suited than the other predictors for more conservative recommendation systems. This is also confirmed by the analysis of the next section, which looks at the probability that the user makes the wrong decision based on a tuneable probability threshold.

5.4 Computing the Probability of Making Wrong Decisions

In the previous section, we showed that the user's utility function $U(\cdot, \cdot)$ induces a unique threshold p^* that completely defines her "Go/No go" policy, namely:

$$\begin{cases} \text{Go} & \text{if } \mathbb{P}(\text{OK}) \geq p^* \\ \text{No go} & \text{otherwise} \end{cases}$$

In this section, we study the reverse question: given a threshold, what is the proportion of the time that a user decides to go to a station but does not find a bike: $(\text{Go}, \emptyset)$; and what is the proportion of the time that a user decides not to go to a station while there was in fact a bike available: $(\text{No go}, \text{OK})$. For instance, the policy (*i.e.* threshold) $p^* = 0$ corresponds to the "always go" behavior, whereas a policy $p^* = 1$ will always lead to a "No go" decision. We are now interested in how often the different predictors are misleading from the user's perspective, depending on her threshold policy.



(a) Probability of $(\text{No go}, \text{OK})$

(b) Probability of $(\text{Go}, \emptyset)$

Figure 10: False positive and false negative probabilities as functions of the threshold probability p^* for a prediction horizon of $h = 30$min (lower is better).

In Figure 10 we plot the proportion of bad decisions (wrong "go" decisions or wrong "no go" decisions) for a prediction horizon of 30 minutes as a function of the parameter p^*. When this threshold is set to 0, the proportion of wrong "go" decisions is 11%. This proportion corresponds to the proportion of wrong decisions of an "always go" policy. It is consistent with the value 89% of Figure 9a. For both HP and QMP, the proportion of wrong "go" decisions decreases as the threshold p^* increases while the proportion of wrong "no go" decisions increases. This is due to the fact that, as p^*

increases, the user less often decides to go to the station and therefore misses more opportunities. The situation is different for LVP: In this case, for any threshold $p^* \in (0, 1)$, the decisions taken by a user using LVP are policy-independent since LVP returns a deterministic value (LVP returns either "the trip is feasible with probability 1" or "the trip is feasible with probability 0"). This shows an advantage of informing the user of the probability of the trip feasibility rather than just a recommendation to go/not go: A user can set her own risk-level and decide to go as a function of this level.

6. CONCLUSIONS

We presented an approach to make forecasts of availability in bike sharing systems (BSS). Our approach uses a queuing theoretic model. Unlike previous techniques based on time-series analysis, our model is explanatory in the sense that it provides an explicit (time-inhomogeneous) dynamics, which is successfully validated across an entire one-year dataset of all stations' activity in *Vélib'*, the BSS of the City of Paris.

The model naturally allows one to make *probabilistic* forecasts, *i.e.*, predictions of the probability distribution of the state of a station, whereas previous work has focused on point estimators only. Shifting to this setting requires us to revisit the notion of accuracy of a predictor, which until now was measured by the root mean square error between the estimate and the actual observation. In this paper, we challenged this view and proposed instead new scoring rules for BSS predictors.

We believe that the possibility of making probabilistic forecasts has significant added value, since it may broaden the scope of the applicability of predictive models. For instance, it more directly provides user-centric quantities of interest, useful for journey planning, such as the probability of finding a bike at the origin station (and dually, of finding an empty slot at the destination station). Furthermore, it can be used in more sophisticated recommendation systems that rank stations not only on their expected number of available bikes, but also on higher-order moments (e.g., the standard deviation), to favour those that exhibit less variability.

The extensive model validation conducted in this paper suggests that our model can in general be used in these applications with high accuracy for prediction horizons up to 2-3 hours. For longer time horizons, we see on average no significant gain of a model that takes into account the current number of bikes parked at a station compared to just using the historical trend. In future work, we plan to examine the prediction accuracy in specific "critical" circumstances, for instance during peak times in popular stations that tend to be either completely empty or full.

A limitation of our model is the implicit assumption of independence which allowed us to consider a station as an isolated queuing system. Future work will aim to relax this, by incorporating the BSS network structure in order to improve the prediction accuracy. Furthermore, we will test the generality of the approach by validating the model against other public datasets of availability in BSS.

7. ACKNOWLEDGMENTS

This work is supported by the EU project QUANTICOL, 600708. The authors would also like to thank JCDecaux for providing the access to the data of the *Vélib'* system.

8. REFERENCES

[1] MoReBikeS: Model reuse with bike rental station data. `http://reframe-d2k.org/Challenge` and `http://www.ecmlpkdd2015.org/discovery-challenges`. Accessed 17/04/2015.

[2] G. Bolch, S. Greiner, H. de Meer, and K. Trivedi. *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. Wiley, 2005.

[3] P. Borgnat, P. Abry, C. Robardet, J.-B. Rouquier, and E. Fleury. Shared bicycles in a city: A signal processing and data analysis perspective. *Advances in Complex Systems*, 14(3):415–438, 2011.

[4] P. DeMaio. Bike-sharing: Its history, models of provision, and future. In *Velo-city Conference*, 2009.

[5] E. Fishman. Bikeshare: A review of recent literature. *Transport Reviews*, pages 1–22, 2015.

[6] I. Frade and A. Ribeiro. Bicycle sharing systems demand. *Transportation: Can We Do More with Less Resources? - 16th Meeting of the Euro Working Group on Transportation - Porto 2013*, 111:518–527, 2014.

[7] C. Fricker and N. Gast. Incentives and redistribution in homogenous bike-sharing systems with stations of finite capacity. *EURO J. Transp. Logist.*, 2014.

[8] C. Fricker, N. Gast, and A. Mohamed. Mean field analysis for inhomogeneous bike sharing systems. In *Aofa 2012, International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms*, 2012.

[9] J. Froehlich, J. Neumann, and N. Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *IJCAI*, volume 9, pages 1420–1426, 2009.

[10] D. George and C. Xia. Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *European Journal of Operational Research*, 211(1):198–207, 2011.

[11] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[12] M. C. Guenther and J. T. Bradley. Journey data based arrival forecasting for bicycle hire schemes. In *Analytical and Stochastic Modeling Techniques and Applications*, pages 214–231. Springer, 2013.

[13] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455–466, 2010.

[14] R. Nair, E. Miller-Hooks, R. Hampshire, and A. Bušić. Large-scale vehicular sharing systems: analysis of Vélib'. *International Journal of Sustainable Transportation*, 7(1):85–106, 2013.

[15] J. W. Yoon, F. Pinelli, and F. Calabrese. Cityride: a predictive bike sharing journey advisor. In *Mobile Data Management (MDM), 2012 IEEE 13th International Conference on*, pages 306–311. IEEE, 2012.